Algorithmic content moderation: Technical and political challenges in the automation of platform governance

Year: 2020 | Citations: 663 | Authors: Robert Gorwa, Reuben Binns, Christian Katzenbach

Abstract

As government pressure on major technology companies builds, both firms and legislators are searching for technical solutions to difficult platform governance puzzles such as hate speech and misinformation. Automated hash-matching and predictive machine learning tools – what we define here as algorithmic moderation systems – are increasingly being deployed to conduct content moderation at scale by major platforms for user-generated content such as Facebook, YouTube and Twitter. This article provides an accessible technical primer on how algorithmic moderation works; examines some of the existing automated tools used by major platforms to handle copyright infringement, terrorism and toxic speech; and identifies key political and ethical issues for these systems as the reliance on them grows. Recent events suggest that algorithmic moderation has become necessary to manage growing public expectations for increased platform responsibility, safety and security on the global stage; however, as we demonstrate, these systems remain opaque, unaccountable and poorly understood. Despite the potential promise of algorithms or 'AI', we show that even 'well optimized' moderation systems could exacerbate, rather than relieve, many existing problems with content policy as enacted by platforms for three main reasons: automated moderation threatens to (a) further increase opacity, making a famously non-transparent set of practices even more difficult to understand or audit, (b) further complicate outstanding issues of fairness and justice in large-scale sociotechnical systems and (c) re-obscure the fundamentally political nature of speech decisions being executed at scale.